U.S. DEPARTMENT OF COMMERCE
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
NATIONAL WEATHER SERVICE
SYSTEMS DEVELOPMENT OFFICE
TECHNIQUES DEVELOPMENT LABORATORY

TDL OFFICE NOTE 83-10

A COMPARATIVE VERIFICATION OF GEM AND MOS

Thomas J. Perrone and Robert G. Miller

July 1983

A COMPARATIVE VERIFICATION OF GEM AND MOS

Thomas J. Perrone and Robert G. Miller

1. INTRODUCTION

GEM is an acronym for a statistical weather forecasting technique which predicts the probability distribution of all surface weather elements hour by hour. GEM uses only the current local surface weather conditions as predictors. Climatological information is also used, in two ways: implicitly, through the Regression Estimation of Event Probabilities (REEP) (see Miller, 1964) in the GEM process, and explicitly, to supply location-specific information to the GEM forecast. From the distribution of probabilities of the forecasted weather events, GEM also makes categorical predictions.

"G" indicates generalized. The same statistical equations can be applied at any location and for any time period. "E" stands for equivalent, because of GEM's equivalence (as a linear approximation) to a Markov chain. "E" also stands for exponential, a characteristic of the particular form of the Markov process necessary to model events which occur in continuous time. "M" indicates that the technique is a Markov process.

An excellent definition of a Markov process as applied to a physical situation, such as weather forecasting, is given by Feller (1950):

> "In stochastic processes the future is never uniquely determined, but we have at least probability relations enabling us to make predictions.... The term 'Markov process' is applied to a very large and important class of stochastic processes.... Conceptually, a Markov process is the probabilistic analogue of the processes of classical mechanics, where the future development is completely determined by the present state and is independent of the way in which the present state has developed...in contrast to processes...where the whole past history of the system influences its future."

The motivation for GEM's development is the need to provide accurate, yet computationally feasible, computer-generated short-term weather forecasting guidance based on the very latest weather information. In general, persistence, though essentially a "no-skill" technique, has been the most skillful guidance available for forecasts of most weather elements for projections ranging from 0- to 6-hours.

Model Output Statistics (MOS) (see Glahn and Lowry, 1972) is now widely accepted as a highly-skilled purveyor of statistical-dynamic weather forecasting guidance. The input to MOS requires data from models, which, however, results in a gap of about 5 hours between upper air observations (about 2 hours from surface observations) and the availability of MOS. The gap results from a combination of two factors: first, the amount of centralized computer time necessary to generate the model output and, in turn, the MOS forecasts; and second, dynamic model instability within the first twelve hours which renders somewhat dubious much of the model output valid for the first 6 hours of the model run.

discusses three miscellaneous topics that were subjects of ancillary statistical experiments. Among the topics discussed are:

a. Use of a collection of multivariate statistical techniques, somewhat related to one another, to make categorical forecasts.

b. Use of blends of local monthly climatology and local hourly climatology to attempt to account more fully for local station effects than by use of local hourly climatology alone.

c. Use of a variation of GEM's P-star process, termed "unaccumulated P-stars", to make categorical forecasts.

Section 7 ends the body of this report with conclusions and some relevant remarks. Terms such as "P-star" have special meaning with respect to GEM. Definitions of such terms, unique to GEM, or which have special usage, are given in Appendix A. In meteorological verification, a number of standard scores are routinely used to report verification results. Among those used in this study are the Heidke Skill Score, percent correctly forecast, threat score (for ceiling and visibility) a chi-square goodness-of-fit measure on margins, mean absolute error (for temperature), mean algebaric error (also for temperature), number of "large" errors (also for temperature), and the Brier Score (for ceiling, visibility, and total cloud amount). These scores are also defined in Appendix A.

2. SCOPE OF STUDY

This section presents background information necessary for understanding the MOS-GEM comparative verification presented in Section 3. It also discusses the scope of the study.

The MOS forecast system has undergone extensive development. Verification of MOS forecasts against both observed conditions and on-station forecaster performance has also been thoroughly documented (see, e.g. Carter, Bocchieri, Dallavalle (1982)). The purpose of the next section is to present comprehensive verification results of GEM against the known guidance standard, MOS.

The general guidelines for the study were to verify weather elements common to both GEM and MOS, for categories of these elements compatible with the two systems, and for projections when the products of both forecast systems were available. The weather elements in common are ceiling, visibility, total cloud amount, temperature, dew-point depression, and wind.

Here are the ways each element was verified:

a. Ceiling. Ceiling was verified as a categorical weather element in six categories. The definition of each category is given in Table 1.

   Scores for the ceiling comprise percent correctly forecast, Heidke Skill Score, a chi-square goodness-of-fit measure on marginals, threat score on the lowest two categories combined, and the Brier Score calculated from the probabilities associated with each ceiling category (see Appendix A for definitions of these scores).

GEM needs only an observation as input, and can make a forecast for any projection, availability of GEM forecasts were dependent only on the availability of initial observations. Within the general scope of the study, MOS archived forecasts were available for projections of 6-, 9-, 12-, 15-, 18-, 21-, 24-, 27-, and 30-hours for temperature and dewpoint depression, and at 6-, 12-, 18-, 24-, and 30-hours for the other elements. Accordingly, verification projections picked for the study were those for which MOS data were available.

Some classification terminology about forecast projections are necessary because MOS and GEM define projections differently. In MOS, the projections are reckoned from the time of the model run which produces the model output predictors (00 GMT or 12 GMT); the observations used by MOS as predictors are usually available 3 hours later (03 GMT or 15 GMT). Sometimes 02 GMT or 14 GMT observations are used, but when no observations are available, MOS uses "backup" equations, which use only model output as predictors. No attempt was made in this study to differentiate among MOS forecasts made with 03 GMT (15 GMT), 02 GMT (14 GMT), or no observational predictors. We took MOS as we found it in the archives, much as it would be available in a real-time setting.

GEM projections are reckoned from the time of the observation used as a predictor. In this study, forecast projections are defined as they are used with GEM. GEM and MOS were comparatively verified in three modes: scientific, operational, and special operational.

In the scientific mode, MOS and GEM share the same observation as input. In the operational mode, the GEM observation used is 6 hours later than the observation used by MOS. In the special observational mode, the GEM observation used is 12 hours later than that used by MOS. The scientific and operational modes are illustrated in the "time lines" of Fig. 2. This figure shows, for these modes, the relative times of the dynamic model run (labeled LFM, for Limited Fine Mesh model), MOS and GEM observation times, and the verification times for 3- and 9-h projections.

The motivation for employing three projection modes was to fully test the validity of GEM from differing viewpoints:

a. The scientific comparison, as its name implies, is a "pure" comparison of GEM and MOS as statistical forecasting techniques. It measures the extent to which GEM, a "classical" (i.e. non-dynamic) statistical technique, which is only limitedly station-specific, can compete against MOS, which not only has model predictors as input but is developed in 6-month seasons for individual stations or small sections of the country.

b. The operational comparison tests GEM's capabilities to operate with later data than centrally-produced MOS and helps to evaluate GEM's usefulness for aviation FT preparation. In this study, GEM forecasts using observations at 09 GMT and 21 GMT simulates production of forecast guidance for 0940 GMT and 2140 GMT FT file times.

c. The special operational comparison tests GEM's capabilities vs MOS's during the periods 00-04 GMT and 12-16 GMT, when the only MOS guidance available is that derived from the previous model run cycle. In this

Score, Heidke Skill Score, and threat score. Under the scientific comparison, GEM in the aggregate is slightly superior to MOS in percent correct, but MOS is favored for the remainder of the scores (See Table 7).

Among the stratified results, the same conclusions hold as for the aggregated results, except in the scientific comparison: GEM and MOS tie in Brier Score for the warm and cool season/03 GMT GEM input time stratification, GEM out-performs MOS in percent correct for the 03 GMT GEM input time stratifications (regardless of season), and GEM has a better threat score for the warm season while MOS is better for the cool season for 03 GMT GEM input time stratification. MOS is better on all scores for 15 GMT GEM input time stratification. MOS chi-squares are preferred over GEM throughout Table 7.

For ceiling at a 9-h projection, for both the operational or scientific comparsion, MOS is favored over GEM in the aggregate for all scores except under the operational comparison for the threat score, where GEM is superior (see Table 8). Among the stratified results, MOS outperforms GEM on all stratifications, except that GEM is superior on the threat score for all operational comparisons and for the cool season/15 GMT GEM input time stratification under the scientific comparison. GEM also outperforms MOS in percent correct for the warm season/21 GMT GEM input stratification under the operational comparison, and for the warm season/03 GMT GEM input stratification under the scientific comparison. MOS chi-squares are smaller than for GEM throughout Table 8, except for the warm season/15 GMT GEM input time stratification under the scientific comparison. For the chi-square measure, smaller is better.

Among the special operational comparisons for ceiling, in the aggregate, GEM is superior to MOS for the 3-h projection; MOS is favored for the 9-h projection except for the threat score, where GEM is superior (see Table 9). Among the stratified results, the same conclusions hold, except for the 9-h projection: GEM is superior in percent correct for the warm season/03 GMT GEM input time stratification, and the two forecast processes tie in percent correct for the warm season/15 GMT GEM input time. Also in the 9-h projection stratification results, GEM's threat score is better in the cool season (regardless of GEM input time), while MOS's threat score is better in the warm season. MOS chi-squares are smaller than for GEM throughout Table 9, except for the cool season/03 GMT GEM input time and warm season/15 GMT GEM input time stratifications for the 9-h projection.

For the operationally critical ceiling categories 1 and 2 (ceilings less than 500 ft), GEM at 3 hours, under the operational comparison, produces 108 "hits" (number correct) for 240 forecasts in the two lowest ceiling categories, while MOS achieves 55 "hits" for 342 forecasts. GEM, therefore, achieves 53 more hits with 102 fewer forecasts than MOS. At 9 hours, also under the operational comparison, GEM produces 38 "hits" with 251 forecasts, while MOS produces 37 "hits" with 219 forecasts. At 9 hours, GEM achieves only one more hit at the cost of 42 additional forecasts than MOS (see Table 10).

In addition to the scores already discussed, Table 10 also displays the biases for each ceiling category. (For definition of bias, see Appendix A.) For the lowest two categories of ceiling, the GEM biases are below one, while MOS is above one, for the 3-h projection. For the 9-h projection, the biases

projection, MOS, in the aggregate, is favored on all measures except the threat score, in which GEM is superior.

For the 3-h projection the aggregated outcomes are also true for the stratifications, except for a no-skill threat score tie between GEM and MOS for the warm season/15 GMT GEM input stratification. For the 9-h projection, the aggregated outcomes hold for the stratifications with these exceptions: GEM is favored with a higher percent correct than MOS in the warm season/15 GMT GEM input stratification, and there is a no-skill tie between GEM and MOS in the warm season/15 GMT GEM input stratification for threat score--otherwise MOS is superior. MOS chi-squares are smaller throughout Table 13.

For the operationally critical visibility categories 1 and 2 (visibilities less than 3 miles), GEM at 3 hours, under the operational comparison, produces 63 "hits" (number correct) for 151 forecasts of categories 1 and 2, while MOS achieves 35 "hits" for 268 forecasts. GEM, therefore, achieves 28 more hits with 117 fewer forecasts than MOS. At 9 hours, also under the operational comparison, GEM produces 23 "hits" with 158 forecasts, while MOS produces 30 "hits" with 160 forecasts. At 9 hours, GEM produces 7 fewer "hits" with 2 fewer forecasts than MOS (see Table 14). Table 14 also shows the biases for each visibility category. For the lowest two categories of visibility, the biases for GEM and MOS are greater than one, except for GEM at 3 hours, where the biases of the lowest two categories are below one.

C. Total Cloud Amount

For the element total cloud amount at a 3-h projection under the operational comparison, GEM in the aggregate is superior to MOS on all scores (see Table 15). Under the scientific comparison, the reverse is true.

Among the stratified results, GEM is superior for all stratifications of the operational comparison, except for a tied Brier Score with MOS for the warm season/09 GMT GEM input time stratification, and for some of the chi-square scores.

Under the scientific comparison, MOS is favored in the Brier Score for all stratifications. GEM achieves a higher percent correct for the cool season, MOS in the warm season. MOS is favored in the Heidke Skill Score for all stratifications, except for a GEM-MOS tie in the cool season/15 GMT GEM input time stratification.

Throughout Table 15, GEM chi-squares are larger than those of MOS, except for these stratifications: under the opertional comparison, cool season/21 GMT GEM input time; under the scientific comparison, warm season/03 GMT GEM input and cool season/15 GMT GEM input stratifications.

For a 9-h projection under both the operational and scientific comparisons, MOS is favored over GEM for those scores reported in the aggregate (see Table 16).

Under the operational comparison, among the stratifications, MOS is favored over GEM on all measures, except for certain chi-square measures; a tie in the

season/15 and 21 GMT GEM input time stratification. For the cool season/03 GMT GEM input time stratification, the results are indeterminate, as the scores for each forecast process are of similiar magnitude but of opposite sign.

Throughout Table 18 GEM chi-squares generally are larger than those for MOS, except for the cool season/09 GMT GEM input time stratification under the operational comparison.

For temperature at a 6-h projection, using either the operational or scientific comparisons, MOS is generally favored over GEM for each measure, whether viewed in the aggregate or for each of the stratifications (see Table 19). The only exceptions under both the operational and scientific comparisons, are for the mean algebraic error, in which GEM is superior for the warm season/15 GMT and 21 GMT GEM input time stratifications. Throughout Table 19, MOS chi-squares are smaller than those of GEM.

Turning to the special operational comparison, for the 3-h projection, GEM, in the aggregate, is superior for the mean absolute error, number of large errors, percent correct, and Heidke Skill Score. MOS, however, achieves in the aggregate a smaller mean algebraic error (see Table 20).

Among the stratifications for the 3-h projection, for mean absolute error, number of large errors, and percent correct, GEM is superior to MOS for each stratification except that of the cool season/15 GMT GEM input time. GEM's Heidke Skill Score is superior for the 03 GMT GEM input time stratifications (regardless of season) while MOS is favored for the warm season/15 GMT GEM input time stratification. The magnitude of the MOS mean algebraic error for the cool season stratifications (regardless of GEM input time) is lower than for GEM, while the warm season algebraic error comparisons for the two forecast processes are indeterminate, because the scores are of similar magnitude but opposite sign.

For projections of 3 hours, the GEM chi-square measures are larger than for MOS for each of the stratifications, except for the cool season/03 GMT GEM input time stratification, where the GEM chi-square is smaller by 0.1.

The 3-h stratification results contrast somewhat with the aggregate results: in the aggregate GEM is generally superior to MOS, but MOS's performance is superior to GEM's on all measures for the cool season/15 GMT GEM input time stratification.

For 6-h projections, MOS is superior to GEM on all scores, both in the aggregate and for each stratification. The sole exception is the number of large errors for the cool season/03 GMT GEM input time stratification; GEM achieves one fewer number of large errors.

E. Dewpoint Depression

For the element dewpoint depression at a 3-h projection for the operational comparison, GEM, in the aggregate, is superior to MOS for the percent correct and Heidke Skill Score measures (see Table 21). The same result holds for the stratifications, except that MOS is favored on the percent correct measure for

For the scientific comparison for the 3-h projection, MOS is superior, both in the aggregate and for each of the stratifications, for the percent correct and Heidke Skill Score measures.

Throughout Table 24 MOS chi-squares are smaller than those for GEM. The largest difference between the GEM and MOS chi-square values is a noteworthy 211, which favors MOS and occurs in the cool season/21 GMT stratification.

For wind at a 9-h projection, MOS is superior to GEM on all measures. MOS is favored in both the aggregate and among the stratifications for both the operational and scientific comparisons (See Table 25). Throughout Table 25 MOS chi-squares are very much smaller than for GEM. The largest difference is 767, favoring MOS, which occurs in the cool season/09 GMT GEM input time stratification under the operational comparison.

Turning to the special operational comparison for the 3-h projection, GEM in the aggregate is superior to MOS for the percent correct and Heidke Skill Score measures. This result also holds among the stratifications, except for the warm season/15 GMT GEM input time stratification, in which MOS is favored (see Table 26).

For the 9-h projection, MOS is superior to GEM for both percent correct and Heidke Skill Score in the aggregate as well as among the stratifications.

Throughout Table 26, MOS chi-squares are smaller than for GEM. The biggest noteworthy differences between MOS and GEM chi-square values (each favoring MOS) occur in these 9-h projection stratifications: cool season/15 GMT GEM input time (561), warm season/15 GMT GEM input time (412), and cool season/03 GMT GEM input time (209).

G. Summary

A summary of the salient results of the GEM-MOS comparative verification is displayed in Table 27. The table expresses the aggregated results in a fractional form. The number of scores favoring GEM forms the numerator; the total number of scores used for the particular weather element is the denominator. These "fractions" are displayed for the special operational, scientific, and operational comparisons, and for the two projections of each element. Since the chi-square measure is not available in aggregate form, it does not enter as one of the scores used in the "fractions". Displayed, however, in Table 27 are the number of stratifications for which the GEM chi-squares are less than those of MOS (a minimum of zero, maximum of four). Major differences in the aggregated results and the results among the stratifications are tagged and identified in footnotes to Table 27.

We may summarize the results in words this way: The order of elements listed in Table 27 is the order of greatest to least skill for GEM in comparison to MOS (i.e., ceiling is most skillful, wind, least skillful). A natural dividing point appears in the table between total cloud amount and temperature. The elements seem to fall into two groups comprising ceiling, visibility, and total cloud amount (elements of major interest for aviation forecasting) on the one hand and temperature, dewpoint depression, and wind on the other. The first group (major aviation elements) is marked by substantial

observational predictors, are generally not available until sometime between 04-05 GMT and 16-17 GMT. The MOS-GEM results from the scientific comparisons are, therefore, adjusted to reflect:

a. Non-availability of new MOS guidance between 04-05 GMT and 16-17 GMT.

b. The deterioration of MOS forecasts relative to GEM's, when GEM uses later observations as input.

There is little difference in the results in both tables in the interval 09-16 GMT and 21-04 GMT, because the operational and special operational differences between MOS and GEM are similar.

4. BLENDING GEM AND MOS

This section presents results of a composite forecast system, derived by statistically blending GEM and MOS, and compares the results of the composite system against GEM and MOS singly. Mr. Joseph R. Bocchieri, formerly with Techniques Development Laboratory*, suggested the blending experiment (Bocchieri, 1982). In principle, our blending experiment is similar to one carried out by him for precipitation forecasting (Bocchieri, 1979); the chief difference, aside from the weather elements involved, is his evaluation of a later observation used directly and blended with MOS, and ours of GEM (based on a later observation) blended with MOS.

To blend the two systems, we derived eight multiple regression equations. These equations represent the elements of ceiling and visibility, each for two forecast projections (3 and 9 hours). All of these equations were derived to provide guidance under two situations: when current cycle MOS guidance is available, and when only previous cycle MOS guidance is available. The equations are of the REEP (Regression Estimation of Event Probabilities, see Miller, 1964) form and use as predictors five of the six GEM and MOS probabilities. The probability of one category is omitted because of redundancy, since each of the forecast processes (GEM and MOS) sum to one.

Table 30 presents the coefficients and additive constants of the REEP blending equations for the element ceiling for 3- and 9-h projections. The visibility equation elements have a similar form.

The GEM probabilities derived from GEM 03 GMT input data were paired with the MOS probabilities derived from input data from the previous MOS cycle (using previous 15 GMT observational predictors and previous 12 GMT model predictors), while GEM probabilities derived from GEM 15 GMT input data were paired with MOS probabilities derived from the input data also from the previous MOS cycle (using previous 03 GMT observational predictors and previous 00 GMT model output as predictors). Data so paired were aggregated from both cycles to achieve the REEP blending equations of Table 30. The pairing of GEM and MOS probabilities just described corresponds to the special operaional comparison defined in Section 2 and used in the GEM-MOS comparative verification of Section 3.

*Present affiliation: National Weather Service Forecast Office, Washington, D.C.

These improvements are accompanied in some instances by better forecast balance, and in some instances by worse balance, as indicated by the chi-square measures.

Some caution must be exercised in comparing the results of blending with either unblended GEM or MOS alone. The GEM and MOS results reported in Section 3 represent verification on independent data, while the blending experiment results reported in this section represent regression fits on that same "independent" data. It would be quite reasonable to expect some "shrinkage" of the blending experiment results if verified on real independent data. However, the amount of shrinkage is deemed to be minor since the number of fitted regression coefficients is very small compared to the sample sizes.

Blending of the sort described here would be an interesting medium for combining MOS and GEM. Blending is an implicit way of utilizing the crossover information reported in the previous section of this report to improve upon the results of either product alone. Also, blending obviates conflicts in the guidance offered by the two systems separately. Blended forecasts would be self-consistent from projection to projection.

Some difficulties, though, might arise in implementing blending. To blend MOS into GEM, yet preserve GEM's capability to forecast for any hour, would require separate blending equations for each difference between the time of the observation used as input to GEM and the time of the observational predictors used in MOS. This requirement follows because centralized MOS is fixed in time as it is generated only twice daily, while GEM is not. The amount of computer storage necessary to hold all the blending equations might be so large as to increase beyond acceptable limits GEM's size for mini-computer applications. Consider, though, blending GEM into MOS. If only the blending equations from this section's experiment were used, it would be possible to issue "updated" MOS guidance shortly after 03 GMT (15 GMT) (based on the previous cycle) and after 09 GMT (21 GMT) (based on the current cycle). Fresh guidance at these times appear to be important for aviation support, in view of FT file times (1540 GMT, 0940 GMT, and 2140 GMT in continental U.S. NWS locations).

5. FEEDBACK

Following completion of the GEM-MOS comparative verification, we examined closely the residuals of the GEM and MOS forecasting processes. The term "residual" is commonly used in meteorological statistics to refer to the difference between what was forecast (usually by a regression, or regression-like process, such as underlies both MOS and GEM) and what was actually observed. Contemplation of residuals, to gain insight into the performance of a regression fitting process, is strongly advocated by data analysts such as Tukey (1977) and experts in regression analysis such as Draper and Smith (1981).

We chose for detailed analysis the data set for the element temperature under the scientific comparison for the warm season/15 GMT GEM input time (12 GMT MOS cycle) stratification. Temperature was an element in which GEM performed less well than MOS. GEM results for the warm season/15 GMT stratification, while neither the best nor the worst in comparison to MOS,

determined separately for each station solely from its own data. These mean absolute error results are shown in column seven of Table 33.

Fig. 5 (see Miller, 1981, page 50) demonstrates in graphical form some of the similarities and differences among the three error feedback schemes. Fig. 5a shows the form of the first error feedback scheme, in which all the data from the 21 stations are grouped together and a single regression line has been fitted to these data. Fig. 5b shows the form of the second error feedback scheme, in which a single slope is derived for all stations from all station's data taken together, but a separate intercept is determined for each station. Fig. 5c shows the form of the third error feedback scheme, in which both slope and intercept have been individually fitted to each station's data.

The overall results, summarized by the weighted average of the mean absolute error for columns five, six, and seven of Table 33, indicate decreasing mean absolute error as the regression method becomes more station-specific. The best overall reduction in GEM mean absolute error (column seven of Table 33 compared with column three) is 0.76°F, an improvement of 17.5% over unadorned GEM.

Comparatively for MOS, mean absolute errors (station-by-station and weighted average), for MOS without feedback are displayed in column eight of Table 33. The mean absolute temperature error results obtained by applying to MOS the most station-specific of the three error feedback schemes (as in Fig. 5c) are given in column nine of Table 33. The weighted average reduction in the MOS mean absolute temperature achieved by applying this third error feedback scheme is 0.16°F, a 5.3% improvement.

The results in Table 33 should be viewed with some caution, however. The unadorned GEM and MOS error statistics are derived from independent test data, while the statistics documenting the application of the feedback processes result from dependent data. Consequently, we expect some shrinkage in the improvements resulting from application of feedback to independent data.

The benefits to be obtained from applying feedback are accompanied by some costs. The smallest improvements, resulting from simple feedback of the mean algebraic error, cost the least. To obtain the mean algebraic error, a representative sample of GEM forecasts and verifying observations is needed. The more sophisticated schemes, which employ feedback of the previous day's forecast errors, perform better, but at higher cost. With these schemes, not only must a properly constructed sample be used to derive regression coefficients, but the previous day's GEM forecast and verifying observation must also be available and carried along by the GEM forecast process. The need to carry along this additional information complicates somewhat GEM's straightforwardness as a forecasting procedure.

There appears, however, to be substantial benefit available, at little real additional cost, when error feedback is used with MOS: The MOS temperature improvement is on the order of 5%. The cost of carrying along the previous day's forecast and verification temperature is relatively small when viewed from the perspective of the large, centralized computer environment which produces MOS.

wind, with 21 categories, the element's categorical space contains 21 dimensions, and each of the 21 elements has a point (a centroid) in the 21 dimensonal space.

A GEM wind probability forecast is a single point in the 21 vector space, determined by the forecasted GEM probabilities for each of the 21 categories. The geometric (Euclidian) scheme calculates, in a straightforward geometric sense, the distance between the GEM forecasted point and each of the 21 centroids, and assigns the forecast to the category whose centroid is "closest", in the Euclidian sense, to the point represented by the GEM probability forecast.

Another categorical decision-making scheme employs a refinement of the Euclidian distance concept, using Σ, called Mahalanobis distance. A spin-off from the Mahalanobis-distance scheme of categorical decision-making is a refinement of the probabilities of each element's categories. The refined, or a posteriori probability, is defined as the probability given that the forecast process was employed, and is obtainable from the Mahalanobis distance (for more details, see Miller, 1962, pp 6-9). Multivariate statistical theory suggests that such a posteriori probabilities should be "sharper" (i.e. should produce lower Brier Scores) than the GEM-forecasted probabilities used as input into the process, when there is underlying multivariate normality in the distribution.

Neither the Euclidian nor Mahalanobis distance classification schemes resulted in better categorical forecasts than the extant GEM P-star thresholding process, for the elements of wind or total cloud amount, suggesting a lack of multivariate normality. Also, the a posteriori probabilities resulting from the Mahalanobis-distance procedure were not, as measured by Brier Scores, "sharper" than the input GEM-forecasted probabilities.

We tried weighting the GEM forecast probabilities and the Mahalanobis-distance process a posteriori probabilities together using the weightings in Table 34.

None of the weightings produced either better Brier Scores nor categorically, a larger number of correct forecasts, than achieved by using GEM alone.

We remain optimistic, though, that some improvement in making categorical forecasts may follow from application of multivariate statistical principles, and we continue our search for and evaluation of these kinds of categorical decision-making techniques. We feel that GEM's respectable forecasting ability, reflected in its Brier Scores, suggest a corresponding potential for categorical forecasting improvement.

B. Station-Specific Monthly Climatological Corrections

To more fully account in the GEM forecast process for local station effects, we tried to blend together adjustments for local monthly climatology with adjustments for local hourly climatology. GEM, as comparatively verified against MOS in this study (see section 3), contains an adjustment to the forecast probabilities for the effects of station-specific (local) hourly climatology (for more details see Miller, 1981, pg 77-79). In particular,

forecasted probability exceeds its unaccumulated P-star threshold probability. More than one category may have its forecast probability exceed its P-star, however. A logical extension of the decision rule, then, is to pick the category whose forecast probability most exceeds the category's unaccumulated P-star threshold probability. And, in the event no single category's forecast probability exceeds the category's threshold unaccumulated P-star, pick the category whose probability lies closest to its threshold P-star.

Restated, the unaccumulated P-star process decision rule is:

a. Pick the category whose forecasted probability exceeds its P-star by the largest amount.

b. If no category's forecasted probability exceeds its P-star, pick the category whose forecasted probability is closest to its P-star.

When applied to the data samples used in the GEM-MOS verification, here are the results:

For 3-h projections, whether under the operational or scientific comparison, for all measures (except the Heidke Skill Score and percent correct for the warm season/21 GMT GEM input time stratification, under the operational comparison), the unaccumulated P-star method ("new" method) is better than maximum probability method ("old" method) (see Table 36).

For 9-h projections, the results are mixed, but for every stratification except warm season/09 GMT GEM input time under the operational comparision, the unaccumulated P-star method achieves a lower chi-square value than the maximum probability method, indicating better balance (see Table 37).

The use of unaccumulated P-stars, however, does not change the relative rankings of the forecasting performance between GEM and MOS. Use of the unaccumulated P-star, however, does reduce the chi-square values of GEM when compared with MOS, in some of the stratifications, by rather large amounts.

7. CONCLUSIONS

GEM demonstrates improvements in forecasting skill over MOS, particularly under the special operational and operational comparisons used in this study. This improvement is strongest for the elements most crucial for aviation operational forecasting (major aviation elements): ceiling, visibility, and total cloud amount. GEM's improvement over MOS, though present, is somewhat less pronounced for the remaining lesser aviation elements considered in this study: temperature, dewpoint depression, and wind. As indicated by the crossover tables of Section 3 of this study, for the major aviation elements, the crossover, determined within the resolution of the data used, lies between 5 and 8 hours from the time of the reference input observation. For the lesser aviation elements, the crossover lies between 3 and 5 hours. Based on the results of this study and previous comparisons of GEM with persistence, we conclude that GEM posseses considerable skill of value for short range operational forecasting guidance.

Tukey, J. W., 1977: Exploratory Data Analysis. Addison-Wesley Publishing Co., Reading, Mass, 688 pp.

Whiton, R. C., 1977: Markov processes. Selected Topics on Statistical Meteorology, Ed., R. G. Miller, Air Weather Service Tech. Report AWS-TR-77-273, Scott AFB, Ill., 45 pp. (Available from Air Weather Service Headquarters, Scott AFB, Ill.)

Wiener, N., 1948: Cybernetics. The Technology Press, MIT, John Wiley and Sons, New York, 194 pp.

_____, 1950: Extrapolation, Interpolation, and Smoothing of Stationary Time Series. The Technology Press, MIT, John Wiley and Sons, New York, 163 pp.

_____, 1956: Nonlinear prediction and dynamics. Proceedings of the Third Berkeley Symposium, Ed., J. Neyman, University of California Press, Berkeley, 247-252.

Appendix A
Definitions of Terms

AFOS: Automation of Field Operations and Services. NWS field station computer system which stores and displays centrally-prepared forecast products as well as collectives of weather observations. Allows field forecasters to compose forecast and warning text products and transmit them to users, and posseses limited capability to run on-site applications programs.

AWS: Air Weather Service. Weather forecasting agency of U.S. Air Force.

Backup Equations: MOS forecast equations derived solely with model output parameters as predictors. Used operationally when surface observations are unavailable.

Bias: Equal to the number of forecasts of an event divided by the number of times the event occurred. A bias of 1.0 is perfect, less than one implies underforecasting, greater than one, overforecasting.

Binary Variable: A variable having a value zero or one. Binary variables, such as used in GEM, are also called "dummy" variables.

Blending: A regression technique which uses predicted probabilities produced by both GEM and MOS for all (less one) categories of ceiling as predictors to produce a new refined probability forecast for each ceiling category. A similar procedure is used for visibility.

Brier Score: $[\Sigma_{\text{all events}}(\text{Probability of an event} - (\text{one, if event occured, zero, if it did not}))^2]/(2 \cdot \text{number of cases})$. Lower values are preferred.

Heidke Skill Score: $\Sigma_{\text{all categories}}$[(Hits - Expected Hits due to chance)/(Total number of cases - Expected Hits due to chance)].

Hits: Number of correct forecasts.

LFM: Limited-Area Fine Mesh. A dynamic modeling system which, over the United States and nearby contiguous regions, uses a smaller grid-length ("mesh") than the National Meteorological Center's hemispheric and global models. Some predictors from the LFM are used in MOS.

Mean absolute error: $\Sigma_{\text{all cases}}$[absolute value (Forecast - Observed)]/Number of cases.

Mean algebraic error: $\Sigma_{\text{all cases}}$[(Forecast-observed)]/Number of cases.

MOS: Model Output Statistics. A dynamical-statistical weather forecasting technique which uses model output and surface and upper air observations as predictors.

Number of Large Errors: A count of events where the forecast and observed temperatures differ by 10°F or more.

OBS: Surface Weather Observations.

Operational Comparison: Method of comparing MOS and GEM where different observational predictors are used by GEM and MOS. GEM uses later observational information than MOS, at a time chosen to be approximately one hour before FT file time. For example, MOS uses 03 GMT surface observational parameters while GEM uses those at 09 GMT.

P-star (P*): A probability value, which, if exceeded by the forecast probability, would initiate a categorical forecast of the event.

Percent correctly forecast: The number of "Hits" divided by the total number of cases, expressed as a percentage.

REEP: Regression Estimation of Event Probabilities. A regression technique where the predictands are only zero or one. The objective is to estimate the probability that the event one will occur.

Residuals: The difference between the fit produced by regression on data, and the data values themselves. Analysis of residuals can provide indications of distributional forms and biases in the regression analysis, and can suggest ways to improve regression fit.

Scientific comparison: Method for comparing MOS and GEM where both forecast techniques use the same surface observational parameters as predictors.

Shrinkage: Degradation of forecast performance on independent data when compared with performance on dependent data. Shrinkage is small when the number of cases in the data sets is large compared with the number of terms fitted by the regression.

Table 1. Ceiling category definitions.

Category Number	Category Definition (ft)
1	<200
2	200-400
3	500-900
4	1000-2900
5	3000-7500
6	>7500

Table 2. Visibility category definitions.

Category Number	Category Definition (mi)
1	<1/2
2	1/2-7/8
3	1-2 1/2
4	3-4
5	5-6
6	>6

Table 4. Dewpoint depression category definitions.

Category Number	Category Definition (°F)
1	0
2	1
3	2 - 4
4	5 - 7
5	8 - 11
6	12 - 15
7	16 - 19
8	20 - 25
9	26 - 35
10	36 - 50
11	51 - 99

Table 5. Wind category definitions.

Category	Wind Direction (°) and Speed (kt)
1	Calm or less than 2
2	020 - 050/2-9
3	020 - 050/10-19
4	060 - 100/2-9
5	060 - 100/10-19
6	110 - 140/2-9
7	110 - 140/10-19
8	150 - 190/2-9
9	150 - 190/10-19
10	200 - 230/2-9
11	200 - 230/10-19
12	240 - 280/2-9
13	240 - 280/10-19
14	290 - 320/2-9
15	290 - 320/10-19
16	330 - 010*/2-9
17	330 - 010*/10-19
18	020 - 100/$\geq$20
19	110 - 190/$\geq$20
20	200 - 280/$\geq$20
21	290 - 010*/$\geq$20

* through 360

Table 7. Ceiling comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

Element: Ceiling

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.161	.139	.191	.153	.112	.102	.163	.136	.161	.135*
% Corr	73.1	80.4	68.6	78.6	80.8	85.4	73.6	79.9	73.4	80.7*
Heidke	.380	.487	.394	.556	.391	.419	.418	.500	.396	.497*
Chi Sq	1.98	29.6	3.61	16.9	2.81	30.3	2.82	54.2		
Threat	.184	.271	.241	.404	.120	.250	.140	.321	.201	.345*
Sample Size	2720		3194		2141		3091		11146	

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.108	.108	.142	.142	.128	.133	.158	.161	.136*	.138
% Corr	82.4	85.1	78.3	79.5	81.0	79.8	76.5	75.3	79.3	79.7*
Heidke	.494	.472	.538	.513	.532	.413	.537	.477	.526*	.472
Chi Sq	4.52	25.7	.79	29.9	1.28	50.1	4.36	55.7		
Threat	.350	.360	.366	.354	.208	.159	.362	.322	.351*	.319
Sample Size	2518		3187		2435		3116		11256	

* Signifies superiority (shown only for aggregated)

Table 9. Ceiling comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 9 hours, for 9-h projection.

Element: Ceiling

Special Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.127	.107	.179	.144	.155	.133	.193	.161	.166	.138*
% Corr.	78.4	85.3	70.8	79.3	76.3	79.8	68.5	75.2	73.0	79.6*
Heidke	.368	.463	.382	.511	.381	.413	.374	.474	.377	.469*
Chi Sq.	2.60	27.5	.97	29.6	2.72	48.1	8.48	56.9		
Threat	.060	.386	.174	.362	.077	.182	.111	.319	.126	.325*
Sample Size	2458		3113		2438		3190		11199	

Special Operational Comparisons

Projection: 9 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.167	.179	.205	.212	.115	.124	.172	.185	.168*	.178
% Corr.	72.6	73.8	65.8	63.2	80.1	80.1	72.2	67.7	72.2*	70.5
Heidke	.371	.340	.345	.317	.341	.290	.372	.309	.358*	.314
Chi Sq.	2.45	28.1	7.86	6.25	17.5	4.44	6.42	28.5		
Threat	.132	.122	.170	.219	.024	.000	.096	.143	.137	.169*
Sample Size	2460		3110		2429		3179		11178	

*Signifies superiority

Table 11. Visibility comparative GEM-MOS verification scores. Under the operational comparison GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations. Forecasts are valid 3 hours after GEM input time.

Element: Visibility

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.171	.171	.154	.132	.092	.077	.110	.090	.134	.119*
% Corr.	72.3	76.7	74.4	82.5	85.7	88.6	81.8	86.9	78.1	83.5*
Heidke	.342	.374	.301	.492	.368	.469	.321	.520	.330	.467*
Chi Sq.	6.63	55.3	10.4	14.6	1.49	5.86	3.78	1.01		
Threat	.129	.210	.239	.343	.000	.143	.211	.393	.202	.319*
Sample Size	2753		3208		2153		3129		11243	

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.073	.073	.087	.086	.084	.091	.115	.118	.091*	.093
% Corr.	89.1	90.6	86.2	88.5	87.8	86.7	80.6	80.1	85.6	86.3*
Heidke	.494	.523	.459	.521	.521	.499	.429	.447	.472	.496*
Chi Sq.	.49	3.75	4.31	8.34	5.12	9.02	13.7	14.0		
Threat	.313	.278	.413	.474	.000	.000	.372	.324	.374*	.347
Sample Size	2553		3204		2443		3142		11342	

*Signifies superiority

Table 13. Visibility comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after input time for 3-h projection; 9 hours, for 9-h projection.

Element: Visibility

Special Operational Comparisons

Projection: 3 hours

Score	GEM Input Obs Time 0300GMT: Season Warm MOS	Warm GEM	Cool MOS	Cool GEM	GEM Input Obs Time 1500GMT: Season Warm MOS	Warm GEM	Cool MOS	Cool GEM	Aggregated MOS	Aggregated GEM
Brier	.091	.072	.109	.087	.112	.091	.135	.118	.113	.093*
% Corr.	84.4	90.7	81.0	88.3	82.4	86.7	78.7	80.1	81.4	86.2*
Heidke	.256	.530	.299	.518	.273	.500	.344	.443	.297	.496*
Chi Sq.	4.03	3.96	3.10	7.76	6.59	8.78	6.24	13.9		
Threat	.087	.333	.211	.481	.000	.000	.195	.313	.190	.348*
Sample Size	2472		3150		2465		3197		11284	

Special Operational Comparisons

Projection: 9 hours

Score	GEM Input Obs Time 0300GMT: Season Warm MOS	Warm GEM	Cool MOS	Cool GEM	GEM Input Obs Time 1500GMT: Season Warm MOS	Warm GEM	Cool MOS	Cool GEM	Aggregated MOS	Aggregated GEM
Brier	.182	.203	.164	.176	.096	.099	.114	.117	.139*	.148
% Corr.	69.3	65.3	72.3	63.6	85.5	87.1	82.4	78.8	77.4*	73.4
Heidke	.321	.223	.274	.172	.247	.182	.269	.264	.277*	.211
Chi Sq.	1.29	46.5	9.89	48.2	4.91	31.6	11.3	16.9		
Threat	.099	.115	.169	.181	.000	.000	.060	.148	.131	.156*
Sample Size	2471		3150		2465		3197		11283	

*Signifies superiority

Table 15. Total cloud amount comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

Element: Total Cloud Amount

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.291	.291	.269	.237	.296	.272	.279	.240	.282	.258*
% Corr.	53.7	55.0	57.1	65.8	52.0	58.4	54.6	63.4	54.6	61.1*
Heidke	.367	.390	.407	.508	.337	.442	.388	.492	.378	.462*
Chi Sq.	1.05	78.0	42.8	80.3	5.74	7.99	68.1	4.81		
Sample Size	2735	2713	3208	3169	2153	2151	3129	3123	11225	11156

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.232	.246	.221	.225	.255	.289	.246	.252	.238*	.251
% Corr.	63.3	60.9	66.2	66.4	59.7	56.0	61.1	62.9	62.7*	62.0
Heidke	.468	.449	.514	.505	.445	.412	.482	.482	.480*	.466
Chi Sq.	7.60	6.89	19.2	59.0	20.6	49.6	57.2	13.8		
Sample Size	2536	2532	3204	3187	2443	2439	3142	3122	11325	11280

*Signifies superiority

Note: Sample sizes for GEM and MOS are slightly mismatched because of differences in total cloud amount definitions in the two forecast processes (see Section 2).

Table 17. Total cloud amount comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 9 hours, for 9-h projection.

Element: Total Cloud Amount

Special Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.282	.247	.267	.225	.298	.288	.298	.254	.286	.252*
% Corr.	54.4	60.7	57.2	66.4	53.5	56.3	50.3	62.4	53.8	61.8*
Heidke	.343	.446	.384	.504	.342	.415	.340	.477	.354	.464*
Chi Sq.	5.12	6.41	24.9	54.4	3.62	45.8	62.1	14.3		
Sample Size	2472	2466	3150	3134	2446	2440	3197	3176	11265	11216

Special Operational Comparisons

Projection: 9 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
Brier	.304	.330	.286	.294	.308	.339	.294	.314	.296*	.318
% Corr.	51.2	45.1	52.5	54.9	49.2	41.5	51.4	48.9	51.2*	48.1
Heidke	.335	.263	.345	.349	.290	.220	.345	.284	.331*	.284
Chi Sq.	1.53	29.2	23.9	66.3	2.92	20.7	88.3	62.5		
Sample Size	2471	2451	3150	3112	2446	2443	3197	3190	11264	11196

*Signifies superiority

Table 19. Temperature comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 6 hours after GEM input time.

MABS Error = Mean Absolute Error
MALG Error = Mean Algebraic Error
No. LG Errors = Number of Large Errors ($\geq$ 10°F)

Element: Temperature

Operational Comparisons

Projection: 6 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	2.41	4.22	2.86	3.91	2.78	3.56	2.96	4.39	2.20*	4.05
MALG Error	.23	-1.92	.43	.90	-.55	-.03	-.08	-.32	.32*	.79
No. LG Errors	40	237	48	166	34	73	64	244	186*	720
% Correct	56.1	37.8	48.8	36.7	50.7	40.2	49.0	33.4	51.1*	36.8
Chi Sq.	12.4	44.4	12.1	27.7	9.98	36.5	8.56	27.7		
Heidke	.506	.302	.437	.302	.441	.319	.437	.267	.455*	.296
Sample Size	3031		3250		2381		3225		11887	

Scientific Comparisons

Projection: 6 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
MABS Error	1.97	2.95	2.56	3.44	3.06	4.30	3.11	4.85	2.68*	3.90
MALG Error	.33	.47	.43	-.99	-.53	-.20	-.41	-1.66	.43*	.83
No. LG Errors	9	42	42	112	73	199	104	408	228*	761
% Correct	61.4	46.6	54.3	42.7	47.6	35.2	46.4	31.8	52.3*	39.0
Chi Sq.	5.83	41.8	7.30	16.3	12.9	64.5	11.4	67.3		
Heidke	.563	.392	.495	.367	.412	.268	.415	.253	.470*	.260
Sample Size	2787		3246		2691		3252		11976	

*Signifies superiority

Table 21. Dewpoint depression comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 3 hours after GEM input time.

Element: Dewpoint Depression

Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	42.7	47.3	38.2	46.9	40.4	45.3	37.3	36.9	39.4	43.8*
Chi Sq.	57.1	68.7	103.	84.6	43.3	41.8	53.7	84.3		
Heidke	.272	.327	.227	.338	.312	.377	.272	.282	.267	.327*
Sample Size	2710		3250		2179		3245		11384	

Scientific Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	48.8	47.7	48.1	41.6	48.0	47.5	44.1	37.3	47.1*	43.0
Chi Sq.	18.4	26.6	26.9	42.8	30.7	68.6	50.2	169.		
Heidke	.378	.365	.373	.292	.393	.382	.357	.274	.374*	.322
Sample Size	2473		3246		2467		3252		11438	

*Signifies superiority

Table 23. Dewpoint depression comparative GEM-MOS verification scores. Under the special operational comparison, GEM uses 03 GMT and 15 GMT observations as input; MOS uses 15 GMT and 03 GMT observations, respectively, from the previous cycle. Forecasts are valid 3 hours after GEM input time for 3-h projection; 6 hours for 6-h projection.

Element: Dewpoint Depression

Special Operational Comparisons

Projection: 3 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	37.6	47.3	33.2	41.6	38.9	47.4	33.7	37.2	35.5	42.8*
Chi Sq.	43.4	26.0	111.	42.9	82.6	67.0	88.0	169.		
Heidke	.230	.361	.181	.293	.282	.381	.236	.273	.229	.321*
Sample Size	2495		3244		2446		3241		11426	

Special Operational Comparisons

Projection: 6 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	36.5	36.6	34.0	33.0	39.4	38.5	35.1	31.5	36.0*	34.5
Chi Sq.	75.0	73.6	145.	64.4	79.9	119.	101.	203.		
Heidke	.196	.219	.180	.177	.219	.272	.249	.203	.227*	.214
Sample Size	2495		3244		2446		3241		11426	

*Signifies superiority

Table 25. Wind comparative GEM-MOS verification scores. Under the operational comparison, GEM uses 09 GMT and 21 GMT observations as input; MOS uses 03 GMT and 15 GMT observations. Under the scientific comparison, GEM and MOS both use 03 GMT and 15 GMT observations as input. Forecasts are valid 9 hours after GEM input time.

Element: Wind

Operational Comparisons

Projection: 9 hours	GEM Input Obs Time 0900GMT				GEM Input Obs Time 2100GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	29.3	19.5	30.0	17.9	33.4	23.2	31.5	18.8	30.9*	19.6
Chi Sq.	24.5	554.	24.6	792.	40.3	225.	36.9	335.		
Heidke	.237	.141	.252	.124	.265	.147	.255	.117	.252*	.131
Sample Size	3031		3250		2380		3226		11887	

Scientific Comparisons

Projection: 9 hours	GEM Input Obs Time 0300GMT				GEM Input Obs Time 1500GMT					
	Season				Season					
	Warm		Cool		Warm		Cool		Aggregated	
Score	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM	MOS	GEM
% Correct	36.1	24.0	32.7	23.6	33.0	24.4	33.6	18.2	33.8*	22.4
Chi Sq.	26.9	141.	30.2	244.	26.4	430.	42.2	600.		
Heidke	.294	.160	.268	.167	.278	.178	.284	.123	.280*	.156
Sample Size	2787		3246		2691		3252		11976	

*Signifies superiority

Table 27. Summary of GEM-MOS Comparative Verification Results. Non-chi-square scores are expressed as fractions: numerator is the number of scores for which GEM is superior to MOS; denominator is the number of non-chi-square scores for an element. Chi-square scores are expressed as the number of stratifications (out of 4) for which the chi-square score for GEM is smaller than for MOS.

Element	Ceiling			Visibility			Total Cloud Amount		
Mode	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional
3-h Proj. Scores	4/4	1/4	4/4	4/4	2/4	4/4	3/3	0/3	3/3
# of Chi Sqs. Favoring GEM	0	0	0	1	0	1	1	2	1
9-h Proj. Scores	1/4	0/4	1/4	1/4	0/4	0/4	0/3	0/3	0/3
# of Chi Sqs. Favoring GEM	2	1	0	0	0	0	1	1	1

Element	Temperature			Dewpoint Depression			Wind		
Mode	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional	Special Operat.	Scien-tific	Opera-tional
3-h Proj. Scores	4/5[A]	0/5	4/5[B]	2/2	0/2	2/2	2/2[E]	0/2	2/2[F]
# of Chi Sqs. Favoring GEM	1	0	1	3	0	2	0	0	0
6 or 9-h Proj. Scores	0/5	0/5	0/5	0/2[C]	0/2	0/2	0/2	0/2	0/2
# of Chi Sqs. Favoring GEM	0	0	0	2	1	0[D]	0	0	0[D]

Footnotes:

A - In stratifications, MOS favored over GEM in cool season/15 GMT GEM input time (12 GMT MOS cycle) stratification.

B - In one or more stratifications, MOS favored in the 12 GMT cycle (21 GMT GEM input time) stratifications.

C - GEM superior for warm season/03 GMT GEM input time (previous 12 GMT MOS cycle) stratification.

D - GEM chi-square values very much larger than MOS for 5 stratifications out of 24, for both dewpoint depression and wind.

E - MOS favored on warm season/15 GMT GEM input (previous 00 GMT MOS cycle) stratification.

F - MOS favored on 21 GMT GEM input time (12 GMT MOS cycle) stratifications for both seasons. GEM favored on 03 GMT GEM input time (00 GMT MOS cycle) stratifications for both seasons.

Table 30. Ceiling regression equations blending MOS and GEM. GEM uses 03 GMT and 15 GMT ceiling probabilities; MOS uses 15 GMT and 03 GMT probabilities, respectively, from the previous cycle.

ELEMENT: Ceiling
PROJECTION: 3 Hours

		Predictand Categories					
Predictors		1	2	3	4	5	6
Additive Constant		.001	-.001	.004	-.010	-.036	1.042
MOS Probability for Predictor Categories	1	.238	-.092	.027	.025	.250	-.441
	2	-.040	.335	-.307	-.008	-.080	.100
	3	.011	-.070	.649	-.024	.158	-.728
	4	.016	-.008	-.116	.310	.007	-.207
	5	-.013	.046	.012	-.077	.440	-.406
GEM Probability for Predictor Categories	1	.608	.267	-.082	-.131	-.029	-.634
	2	-.031	.877	.294	-.107	-.098	-.934
	3	-.008	-.060	.721	.078	-.102	-.630
	4	-.005	-.008	.005	.911	-.050	-.853
	5	-.000	-.012	-.027	-.047	.848	-.768

ELEMENT: Ceiling
PROJECTION: 9 Hours

		Predictand Categories					
Predictors		1	2	3	4	5	6
Additive Constant		-.002	-.002	-.004	-.013	-.039	1.061
MOS Probability for Predictor Categories	1	.611	.189	-.033	-.596	-.049	-.122
	2	-.166	.237	-.104	.304	.141	-.412
	3	.024	.233	.814	.081	.195	-1.346
	4	.016	-.062	.022	.604	.029	-.608
	5	-.001	.013	-.062	.042	.836	-.827
GEM Probability for Predictor Categories	1	.623	.371	-.250	-.135	-.491	-.119
	2	.192	.821	.143	-.110	.028	-1.073
	3	-.140	-.141	.602	-.392	-.206	.278
	4	-.007	-.014	-.044	.757	-.030	-.662
	5	.016	-.011	.013	-.158	.479	-.339

Table 31 (continued)

ELEMENT: Visibility
PROJECTION: 3 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.113	.093	.091*	19.5	2.2
% Correct	81.4	86.2	86.4*	6.1	0.2
Heidke	.305	.492	.497*	63.0	1.0
Threat	.190	.351	.355*	86.8	1.1

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	8.12	10.0	3.07*	5.05	7.33

SAMPLE SIZE: 11284

ELEMENT: Visibility
PROJECTION: 9 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.139	.148	.135*	2.9	8.8
% Correct	76.7	72.7	78.5*	2.3	8.0
Heidke	.296	.224	.336*	13.5	50.0
Threat	.131	.157	.173*	32.1	10.2

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	6.88*	26.1	15.0	-8.12	11.1

SAMPLE SIZE: 11283

* Signifies superiority

Table 32 (continued)

ELEMENT: Visibility
PROJECTION: 3 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.134	.119	.114*	14.9	4.2
% Correct	78.1	83.5*	83.4	6.8	-0.1
Heidke	.330	.467	.476*	44.2	1.9
Threat	.202	.319	.325*	60.9	1.9

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	15.3*	45.4	24.2	-8.9	21.2

SAMPLE SIZE: 11243

ELEMENT: Visibility
PROJECTION: 9 Hours

	MOS	GEM	BLENDED	% Improvement OVER MOS	% Improvement OVER GEM
Brier Score	.114	.122	.113*	0.9	7.3
% Correct	81.3	79.9	82.7*	1.7	3.5
Heidke	.301	.238	.327*	8.6	37.4
Threat	.175	.167	.250*	42.9	49.7

	MOS	GEM	BLENDED	Differences (MOS-BLENDED)	Differences (GEM-BLENDED)
Chi Square	8.44*	35.6	19.2	-10.76	16.40

SAMPLE SIZE: 11243

* Signifies superiority
T Signifies Tie

Table 34. Weightings for GEM forecasted probabilities with Mahalanobis-distance *a posteriori* probabilities.

GEM forecasted probabilities	Mahalanobis-distance *a posteriori* probabilities
0.50	0.50
0.90	0.10
0.10	0.90

Table 35. Weightings for station-specific hourly and monthly climatologies.

Station-specific hourly climatology	Station-specific monthly climatology
0.50	0.50
0.67	0.33
0.33	0.67

Table 37. GEM wind comparative verification scores between OLD and NEW categorical selection procedures. Upper half of table displays data for GEM forecasts made from 09 GMT and 21 GMT observations, lower half of table displays data for GEM forecasts made form 03 GMT and 15 GMT observations.

OLD = Maximum probability categorical selection procedure.
NEW = Unaccumulated P-star categorical selection procedure.

ELEMENT: Wind
PROJECTION: 9 Hours

Operational Comparisons

	GEM Input Obs Time 0900 GMT				GEM Input Obs Time 2100 GMT			
	Season				Season			
	Warm		Cool		Warm		Cool	
	OLD	NEW	OLD	NEW	OLD	NEW	OLD	NEW
% Correct	19.5*	18.1	17.9*	16.8	23.2*	22.9	18.8	20.0*
Chi Sq.	554*	561	792	673*	225	140*	335	128*
Heidke	.141*	.129	.124*	.119	.147*	.146	.117	.128*
Sample Size	3031		3250		2380		3226	

Scientific Comparisons

	GEM Input Obs Time 0300 GMT				GEM Input Obs Time 1500 GMT			
	Season				Season			
	Warm		Cool		Warm		Cool	
	OLD	NEW	OLD	NEW	OLD	NEW	OLD	NEW
% Correct	24.0	24.4*	23.6	24.2*	24.4*	23.0	18.2*	17.6
Chi Sq.	141	92*	244	149*	430	275*	600	369*
Heidke	.160	.166*	.167	.174*	.178*	.176	.123	.123
Sample Size	2788		3246		2691		3252	

*Signifies superiority

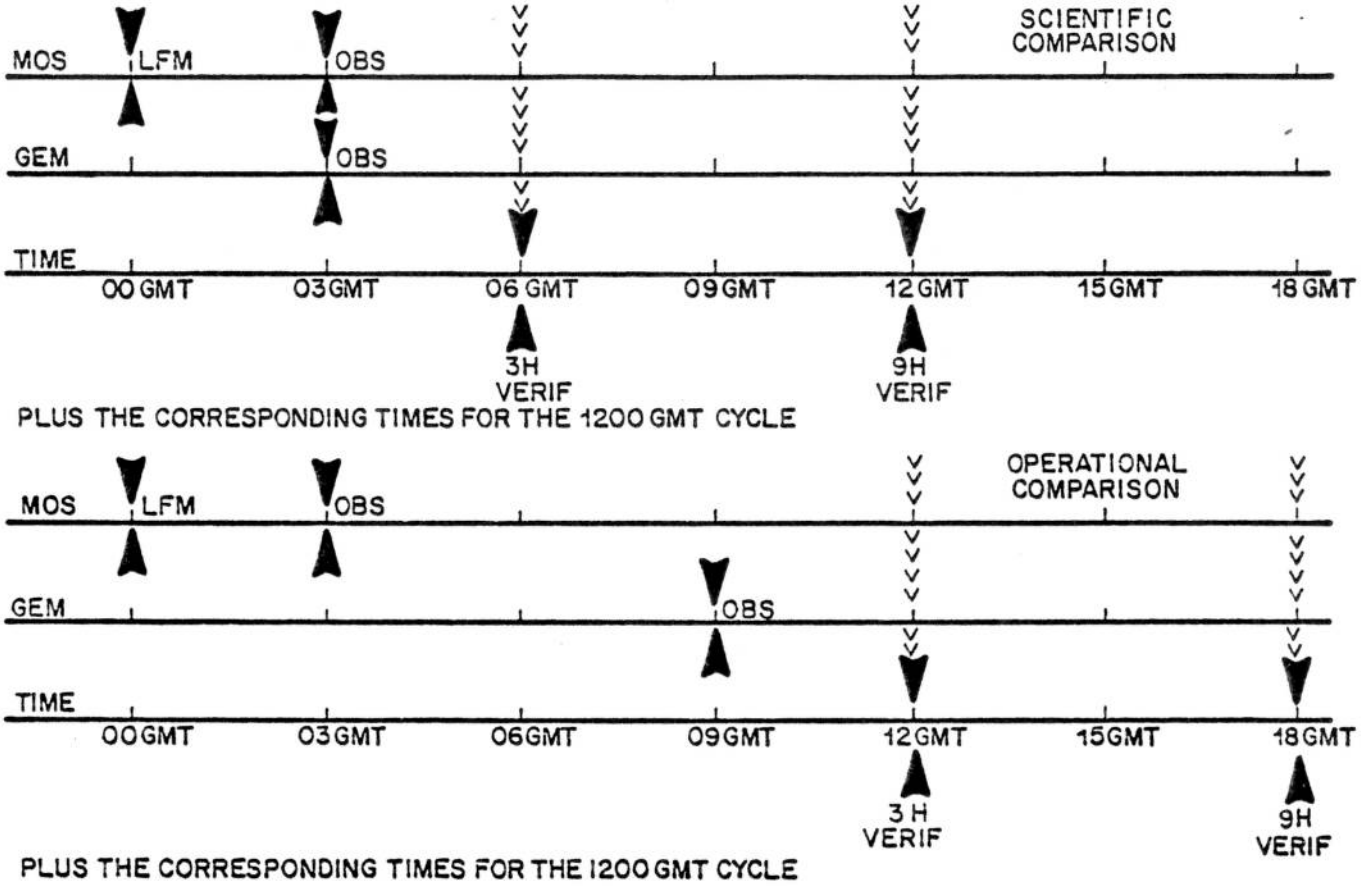

Figure 2. GEM-MOS comparative verification time lines. The upper part of the figure shows the scientific comparison; the lower part shows the operational comparison. Each part indicates the time of the model output (LFM), the surface observations (OBS) and the 3- and 9-h verification times used. The special operational comparison differs from the operational comparison as follows: The MOS and GEM input observation times differ by 12 hours instead of 6, and the MOS LFM predictors came from the previous model cycle, rather than the current cycle.

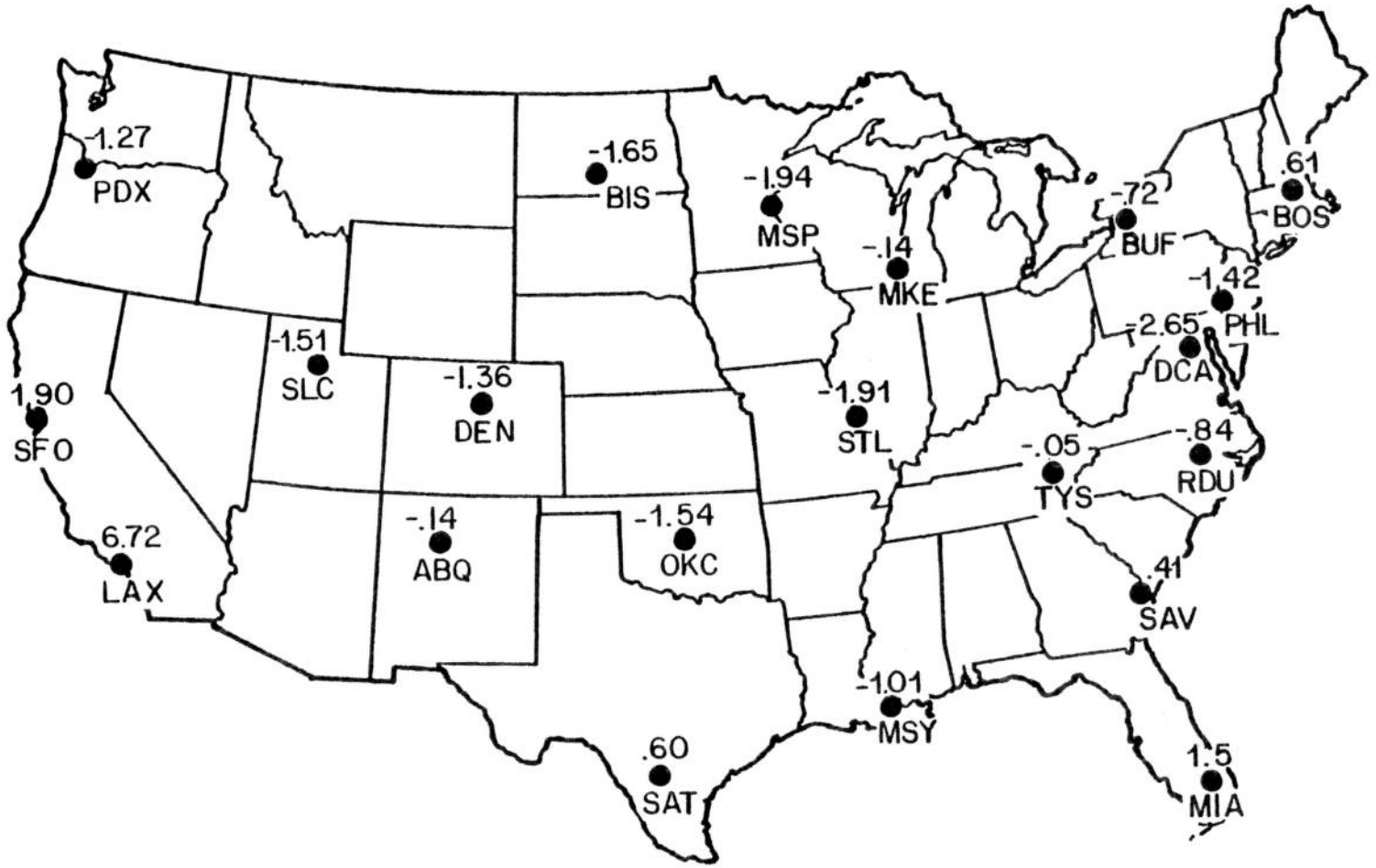

Figure 4. GEM mean algebraic error for 21 stations in the GEM-MOS comparative verification sample.